ARMY RESEARCH LABORATORY

Installation to Production of a Large-Scale General Purpose Graphics Processing Unit (GPGPU) Cluster at the U.S. Army Research Laboratory: Thufir

by Brian J. Henz, John Lazorisak, Jaroslaw Knap, Jason Livingston, and Dale R. Shires

ARL-TR-7085 **September 2014**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-TR-7085 **September 2014**

Installation to Production of a Large-Scale General Purpose Graphics Processing Unit (GPGPU) Cluster at the U.S. Army Research Laboratory: Thufir

Brian J. Henz, Jaroslaw Knap, and Dale R. Shires
Computational and Information Sciences Directorate, ARL

John Lazorisak and Jason Livingston
Lockheed Martin

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2014	Final	June 2011–January 2013

4. TITLE AND SUBTITLE
Installation to Production of a Large-Scale General Purpose Graphics Processing Unit (GPGPU) Cluster at the U.S. Army Research Laboratory: Thufir

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Brian J. Henz, John Lazorisak, Jaroslaw Knap, Jason Livingston, and Dale R. Shires

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CIH-C
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-7085

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
DOD High Performance Computing Modernization Program
10501 Furnace Road
Suite 101
Lorton, VA 22079

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report documents the installation, acceptance testing, vendor interactions, and final system acceptance for production usage of a dedicated high performance computing project investment that was installed at the U.S. Army Research Laboratory facilities on the Aberdeen Proving Ground in MD. The uniqueness of this system lies in the use of the low latency 10 GigE (gigabit Ethernet) network fabric, many-core AMD Opteron central processing units (CPUs), Nvidia Fermi graphics processing unties (GPUs), and a mixture of CPU-only and CPU/GPU nodes. The final accepted production system is stable and capable of nearly 1 PetaFLOPS (10^{12} floating-point operations per second) of single precision performance.

15. SUBJECT TERMS
GPGPU, HPC, hybrid cluster, high performance computing, graphics processing unit

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Brian J. Henz
Unclassified	Unclassified	Unclassified	UU	24	19b. TELEPHONE NUMBER (*Include area code*) 410-278-6531

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

1. Introduction

In February 2012, a dedicated high performance computing project investment (DHPI) was installed at the U.S. Army Research Laboratory (ARL) facilities on the Aberdeen Proving Ground (APG) in MD. This report documents the installation, acceptance testing, vendor interactions, and final system acceptance for production usage. The cluster was unique within the Department of Defense (DOD) at the time due to its use of a low latency 10-gigabit Ethernet (GigE) network fabric, many-core (12) AMD Opteron central processing units (CPUs), Nvidia Fermi graphics processing units (GPUs), and a mixture of CPU-only and CPU/GPU nodes. The combined government/contractor team worked for more than 7 months to complete acceptance of the system and provide feedback to the vendor for system failures and repairs. The final accepted production system is stable and capable of nearly 1 PetaFLOPS (10^{12} floating-point operations per second) of single precision performance.

2. Background

The Thufir system is dedicated to network simulation and emulation with a focus on mobile ad-hoc networks (MANETs) and the Mobile Network Modeling Institute (MNMI). The name Thufir comes from a Mentat character in the Dune book series. Mentats are humans trained as a replacement for computerized calculation (*1*). The MNMI was established in fiscal year 2007 to exploit high performance computing (HPC) through the development of computational software that enables the DOD to design, test, and optimize networks at sufficient levels of fidelity and with sufficient speed to understand the behaviors of network-centric warfare technologies in the full range of conditions in which they will be deployed. Operational goals include the development of scalable computational modeling tools for simulations and emulations, the ability to understand a priori the performance of current and proposed radio waveforms such as the Wideband Networking Waveform (WNW) in the field, and to optimize the network for U.S. Soldiers.

Wireless networks have become a fixture in the modern world, and the shortcomings of such networks are encountered all too often: dropped calls; variations in data traffic bandwidth as a function of distance from a wireless access point or from walls and other objects that block line of sight to an access point; and interference from other radio sources. In everyday life, users often adapt by relocating to a point where there is a better wireless signal, and providers adapt by increasing the density of access points in areas with spotty reception.

The bulk of research conducted in academia focuses on wireless cellular networking as previously described. While aspects of cellular networks carry through to mobile ad hoc networks, the two are distinct in many ways with mobile networks posing greater challenges. In mobile ad hoc networks, access points can move and coverage may vary widely in a region. At times, the access points will cluster together, leaving parts of the map with sparse coverage and parts with compromised service due to competition for available bandwidth as channel subscriptions become saturated. On the battlefield, it may not be an option to relocate receivers to areas with better coverage, which highlights the need to optimize and plan prior to mission as much as possible. Consider the problem of combat in an urban area. Narrow streets and buildings with metal roofs and reinforced concrete walls may interfere with radio reception and access to surveillance data, yet that surveillance information may be key to locating enemy combatants before they inflict casualties on friendly forces.

Mobile networks must be understood at a series of levels, from radio to packet network to communication infrastructure and its resulting impact on the Warfighter. To make this difficult problem tractable, the MNMI has developed a four-pronged approach.

1. The first is MANET simulation, where large-scale HPC assets can be used to test and optimize large radio deployments. These simulations are most often limited to determine whether radios can “see” each other.

2. Second is MANET emulation, where researchers can investigate performance of proposed radios high in the network stack (application layer) all the way to the lower physical layers.

3. Third is MANET experimentation, where live and constructive exercises can capture real radio performance and test interoperability with real and virtual assets. This also results in data sets that can be later mined to fill data gaps and verify models.

4. Fourth is a system that ties together all of these aspects and brings in support for visualization and data analysis. The MNMI and Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance-Network Modernization are addressing all of these topics. Thufir is focused primarily on two efforts, namely simulation and emulation but is also greatly enhancing all four focus areas of the MNMI.

The scale and complexity of mobile ad hoc networks is unique to the DOD, and to the Army in particular, as a mobile fighting force. The military is rapidly becoming a network-centric force, with substantial access to sensor-derived surveillance information as well as an increasingly complicated application layer running over many different devices. This introduces significant advantages to the Warfighter but also brings in new dependencies and new risks from the rapid change in configurations of the MANETs that provide network access across the battlefield.

Unfortunately, it is difficult to design and evaluate mobile ad hoc networks with sufficient fidelity and scale. The research plans and focus of the MNMI are addressing the primary issues associated with optimized MANET planning and execution. Thufir provides dedicated HPC

resources to assist MANET-related research to help researchers identify the principles for MANET optimization and planning without trial-and-error testing that will significantly increase Soldier risk, jeopardize mission success, and not map well to future missions.

3. Dedicated High Performance Computing Project Investment: Thufir

MANET emulations on a large scale, up to 5000 emulated devices, require dedicated resources that allow for the research, development, and evaluation of network algorithms in a controlled environment. In addition, real-time MANET emulated devices may be inserted into live experiments to augment testing and perceived traffic by the physical network devices in the field. This enhances the experience of testers and increases the degrees of freedom that can be evaluated in an experiment. To achieve these goals, a number of technical gaps must be bridged, including the real-time computation of radio frequency (RF) propagation in urban environments and the high-fidelity software emulation of network devices.

MANET emulations require hardware-in-the-loop that can create a controllable, repeatable virtual environment for the testing and evaluation of network devices, both real and emulated. The software emulation of network devices requires low-level access to the Internet Protocol (IP) transport capable network interfaces connected to the host machine. This low-level access allows the emulated network stack to process incoming IP packets or Ethernet frames and forward them through the emulated network stack. Each emulated device requires isolated network access that is provided using virtualization technology. The virtualization technologies include full hardware virtualization such as KVM (kernel-based virtual machine), XEN, and VMware or container-based virtualization such as OpenVZ. Unfortunately, each of these technologies requires a specialized operating system kernel in order to be used. This presents a problem for utilization of shared resources where stability and security are the primary concerns and is a motivating factor to have a dedicated resource such as Thufir.

RF wave propagation models play an essential role in planning, analysis, and optimization of radio networks (*2*). For instance, coverage and interference estimates of network configurations are based on field strength predictions. Approaches for field strength prediction can be divided into semiempirical and ray-optical models. For example, the semiempirical COST-Walfisch-Ikegami model (*3*) estimates the received power predominantly on the basis of frequency and distance to the transmitter. Ray-optical approaches identify ray paths through the scene, based on wave guiding effects like reflection and diffraction. Semiempirical algorithms usually offer fast computation times but suffer from inherent low prediction quality. Ray-optical algorithms feature a higher prediction quality at the cost of higher computation times. For MANET emulation integration, these algorithms must be computed in real-time for each of the propagation paths possible, with an $O(n^2)$ complexity.

GPUs such as those in Thufir have been identified as a solution to provide the raw floating-point performance required to compute each RF propagation path loss in real-time. GPU computing provides great promise for improved performance over traditional CPUs for floating-point intensive applications. The GPU architecture is also ideally suited for accelerating ray-tracing algorithms that are found in the ray-optical approaches for RF wave propagation modeling. In addition, these RF propagation computations must be tightly coupled with the MANET emulation environment to provide real-time (defined to be less than 0.5 s based on routing-table refresh rates) response to computation requests.

The environment currently used within the MNMI for MANET emulations requires that all radio positions be determined a priori because of the costly path loss computations. This means that real-time adjustments based on force-on-force interactions cannot be used in tandem with MANET emulations, thus limiting the utility of the emulation environment. The capability to dynamically alter radio positions and trajectories would add a level of realism to the network emulations and simulations that is not currently possible. In addition to potentially real-time propagation modeling, a GPU-accelerated algorithm would be the basis for adding higher fidelity modeling capabilities such as foliage (*4*) and weather effects.

In general, current commercial simulation software packages only provide low to medium fidelity RF propagation models. High-fidelity models, such as those developed in the electronic battlefield Common High Performance Scalable Software Initiative portfolio (*5*), are required by MANET emulations to provide realistic stimulation to live experiments and for the assessment of emerging technologies. Accurate modeling of RF wave propagation requires path loss computations due to free space distances, wave reflections, refraction, weather effects, and adsorption. Each of these factors contributes to some degree to the computational requirements of the propagation model. For instance, to include foliage in the propagation model, adsorption and scattering by branches and leaves must be considered. Modeling of the forest parameters must be considered, such as tree types (deciduous or coniferous) and tree density (*4*).

Initially, we will be running two software applications with RF propagation models using general purpose graphics processing units (GPGPUs). Both of these applications were developed as part of the MNMI. The first is based on the Irregular Terrain Model (ITM) or Longley-Rice model. The second application is a ray-tracing algorithm using GPUs to compute line-of-sight paths between devices. Ray-tracing is primarily required for urban environments where the reflections and refractions from buildings and walls are of primary importance. Both of the RF propagation models currently execute in parallel on multiple GPUs and were expected to scale well with the number of GPUs.

We have focused primarily on the MANET emulation software extendable mobile ad-hoc network emulator (EMANE) that is capable of scaling up to thousands of MANET devices with the resources provided by Thufir. In addition to an 802.11 a/b/g wireless model and a configurable radio model, a hi-fidelity Soldier Radio Waveform (SRW) model that requires

increased computational effort beyond the standard civilian waveforms is also available. ARL is also investigating the development of additional Army radio waveforms such as WNW and WNaN (Wireless Network after Next) that will have even higher bandwidth requirements than SRW in addition to a potentially increased computational complexity. Emulating a hybrid network consisting of multiple Army radio waveforms will require a network interconnect that can support both high bandwidth (> 1 Gb/s) and low latency (<3 μs) because of the connectedness, $O(n^2)$, of the MANET. This exponential scaling of traffic also exponentially increases the computational requirements of the EMANE software.

The amount of real network traffic generated by MANET emulations presents a unique problem for production machines. Each packet generated from EMANE is broadcast using multicast across-the-network switching fabric. The volume of traffic generated is exponential based on the number of radios and will quickly saturate even a high bandwidth/low latency interconnect such as Infiniband or 10Gb/s Ethernet. This traffic would severely impact the performance of any other parallel application requiring use of the switching fabric, such as the Message Passing Interface (MPI)-based computational mechanics applications. In addition, MNMI research commonly involves investigations of MANET security concerns such as wormhole attacks, denial of service, or steganography. Since a MANET emulation is using hardware in the loop and generating actual network traffic, this research would be indistinguishable from a real attack and can cause potential problems for analysts watching the production networks for suspicious activities.

Running EMANE with any of these virtualization methods requires some advanced networking configurations. This can be achieved through a few methods such as creating a tun/tap interface or bridging an existing interface. A third method is available within EMANE when virtualization is not used. However, this method requires opening a physical interface in promiscuous mode, thus also requiring super-user access. Strict security postures in shared-resource computing systems precluded their use for this type of use.

The preceding discussion provides the basis for the acquisition of Thufir and describes how the system is used to enhance MANET simulation and emulation within ARL. Thufir provides a very low latency 10 GigE network for over-the-air (OTA) communications between radios, real-time RF propagation capabilities through the 456 Nvidia M2070 GPUs, a high speed and capacity storage system from Panasas, and 6576 AMD Interlagos CPU cores for hosting virtual machines (VMs) and running radio models. Other MANET support functions have also been identified, such as large-scale data reduction and mining for support of experimentalists and analysts. The following sections detail the installation, testing, and acceptance phases of Thufir's integration into the MNMI.

3.1 System Specification

- 2.6 GHz 12c Interlagos AMD CPUs
  - 160 compute 2S nodes, 2.67 GB memory/core
  - 114 GPU 2S nodes, 2.67 GB memory/core
  - 24 cores/node and 64 GB memory/node
  - 6576 total compute cores
  - 456 Nvidia M2070 GPUs, 4 GPUs per GPU node, 3 PCIe ×16 bus
- 10 GigE cluster interconnect
  - 10 Gb/s and 40 Gb/s with Gnodal switches
  - 2-stage fat tree
  - Switch port-to-port latency ~282 ns min, ~546 ns max
  - Non-oversubscribed
- (2) 40 TB Panasas PAS12 shelves, dedicated 1 GigE interconnect
  - 2 Login and 2 management nodes, management network

3.2 Installation

The high density of the compute capabilities of GPGPUs also requires facilities with significant power and cooling available. Facilities requirements for a system such as this include raised floors for wiring and cooling with load limits, a reliable source of high-voltage electricity, and significant air- or water-cooling capabilities. For instance, this system requires a maximum of 201 kW of power and 60 ton of cooling. In addition to the supply of power and cooling to this system, the network connectivity is an important issue for remote usage. Remote users from other facilities may require access, which means that these users need clearance checks on record and signed user agreements. The power requirements of up to 201 kW, including power backup using generators, require significant local facilities support. Fortunately, the APG DOD Shared Resource Center site has significant facilities investments and expertise already installed and available for use. The initial installation and power up of the system went smoothly and the DHPI system was co-located with other HPC assets to leverage the currently available facilities.

3.3 Initial Acceptance Testing

As of this writing, in 2014, it is apparent that HPC systems for the foreseeable future are likely to include some type of many-core solution such as the Intel MIC (many integrated core) or GPUs such as the Nvidia Tesla coprocessors. Both of these technologies act as co-processors and therefore add another level of complexity in developing and testing application execution and performance. The acceptance plan developed for this system focused on these nontraditional technologies that are most likely to cause instabilities or have performance problems.

The initial acceptance test suite consisted of a series of tests to check performance and capabilities of the system. These tests were mainly taken from the technology insertion test suite used for acceptance testing of the distributed shared resource center acquisitions. Additional tests included compute unified device architecture (CUDA), OpenCL, and performance of multicast traffic on the network fabric. A 5000-node wireless MANET emulation using EMANE was used to analyze system capabilities to simultaneously support 5000 VMs. The initial tests were performed during the second half of February 2012. These tests were as follows:

1. Demonstrate that the nodes that are equipped with dual redundant power supplies will continue to function in the event one is removed (Twin^2 compute nodes, Admin nodes, login nodes).
2. Demonstrate that the system can be brought to an orderly halt while preserving the file systems.
3. Demonstrate that the system can be brought back up after a full stop and reconnect to the file system.
4. Demonstrate that files can be exchanged between the cluster and another HPC system.
5. Demonstrate that the system supports C, C++, and FORTRAN using the PGI Compiler Suite.
6. Demonstrate that the batch job scheduler, in this case Grid Engine, is operational.
7. Demonstrate that the system supports C, C++, and FORTRAN using the Intel Compiler Suite.
8. Demonstrate that the system supports OpenCL.
9. Demonstrate that the system supports CUDA.
10. Demonstrate that the system supports C and FORTRAN bindings for MPI over the 10 GigE network using OpenMPI or its equivalent.
11. Demonstrate the network latency from different locations within the cluster such as node to node and VM to VM.
12. Demonstrate the network bandwidth.
13. Demonstrate the network multicast scalability.
14. Demonstrate that the system supports IPv4 and IPv6 dual stack functionality.
15. Demonstrate the Panasas features and performance including input/output bandwidth and operation from a VM.

16. Demonstrate GPU performance (NBODY).

17. Demonstrate that the EMANE application can be successfully run on the cluster.

3.4 Stability Issues

Following base performance and capability testing, Thufir began effectiveness testing. Effectiveness testing requires a 97% system availability during a 30-day period with full system utilization, calculated using the following equation:

$$\mathrm{EL} = 100 * \frac{OperationalUseTime(Processor\ hours)}{ScheduledUseTime(Processor\ hours)} \quad (1)$$

where EL is the effectiveness level, or system availability of the complete cluster. In addition to the CPU cores, Thufir includes a number of GPUs that draw the lion's share of power and generate more heat than the rest of the devices in the system. During the run-up to effectiveness testing, we identified a stability issue when running applications that attempted to simultaneously use all of the GPUs, namely four, in a single-system node. Identifying the reason for the instability was not trivial. A more detailed discussion and timeline of the vendor-supplied support is detailed in section 3.6.

The GPUs supported the execution of CUDA and OpenCL applications during the initial acceptance testing. These tests included short computational kernels and ran serially across the GPUs in the system. An OpenCL/MPI application that computes multibody interactions was executed in parallel across all GPUs during effectiveness testing. Initially, the GPUs appeared to fail randomly with the error "NVRM: Xid (0000:07:00): 48, An uncorrectable double bit error has been detected on GPU (00 03 00)" followed by an error stating that the device "fell off the bus." After these errors occurred, the GPUs would be unusable and the system would require a hard reboot in order to return to a usable state. The vendor eventually identified that the issue was related to the 12V power backplane printed circuit board (PCB), figure 1.

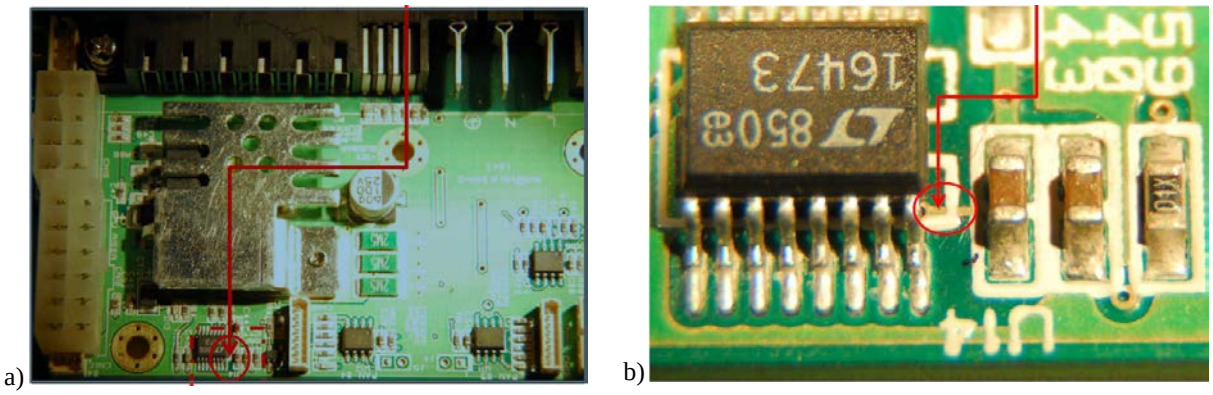

Figure 1. (a) Image showing the trace on the power backplane PCB that needed to be severed. (b) Close-up view of trace to be severed.

Prior to severing the PCB trace, the power backplane was only able to supply 45 A of current to the system, resulting in the GPU instabilities observed. After the trace was severed, the power backplane was able to provide the full 60-A current that the power rail was rated to provide.

3.5 Performance Issues

Multiple performance issues initially plagued the DHPI system as it was spun up for effectiveness testing. These were related to the Interlagos architecture, the MPI implementation, and the 10 GigE network. Eventually, solutions to each of these issues were applied through literature searches, software updates, and system configuration changes.

The first performance issue observed involved job scheduling on the Interlagos many-core processors. Each of the compute and graphics nodes in Thufir contains two 12-core AMD Opteron 6238 processors. This means that if threads are appropriately allocated, the first 12 threads will have dedicated cache and floating point units. When utilizing 13 to 24 cores per node, performance is expected to decrease slightly. With the shared cache, cache misses can be an issue. Initial testing showed the cache misses to increase significantly if more than 12 cores were utilized as shown in figures 2 and 3 for sysctl parameters and parallelization strategy, respectively.

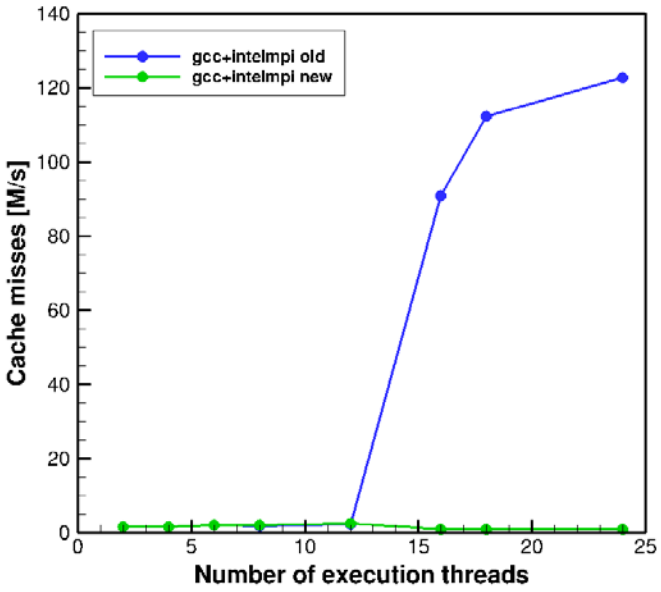

Figure 2. Cache misses vs. number of threads per node showing significant increase in misses when more than 12 threads are executing simultaneously. Old and new refer to before and after setting the sysctl parameter `kernel.randomize_va_space` from 2 (old) to 0 (new).

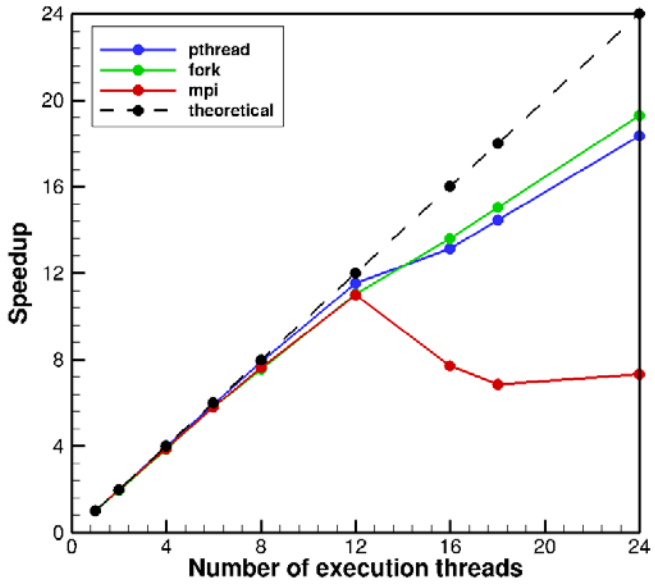

Figure 3. Plot of speedup vs. number of execution threads using pthreads, fork, and MPI. Comparing the performance of the multithreaded applications vs. a tightly coupled MPI-based SIMD application illustrates the effect of L1 instruction cache misses in the MPI library.

A white paper by AMD detailed an issue with instruction-cache cross-invalidations that were the primary cause of these cache misses (*6*). At the hardware level, the first line of the level-1 instruction cache is invalidated when two cache lines with the same physical address are loaded. This can occur when the same code is executing on both cores of a single compute unit in tight loops. This is exactly what occurs in single instruction multiple data (SIMD) applications like those we typically execute, e.g., finite element methods, finite difference, and molecular dynamics.

A temporary solution is to turn off the address-space layout randomization (ASLR) using the following command (*6*):

```
# sysctl –w kernel.randomize_va_space=0
```

After disabling the ASLR and re-running the benchmark tests, we observed a remarkable improvement in performance, indicating that the scaling issue had been identified. Since disabling the ASLR can decrease the system's level of security, a permanent solution was required. The recommended solution was to upgrade the Linux kernel to 3.2-rc1 or higher. We achieved the final solution by updating the entire system from Red Hat Enterprise Linux 6.1 to 6.2 and implementing the kernel patch. Figure 4 shows the marked improvement in performance.

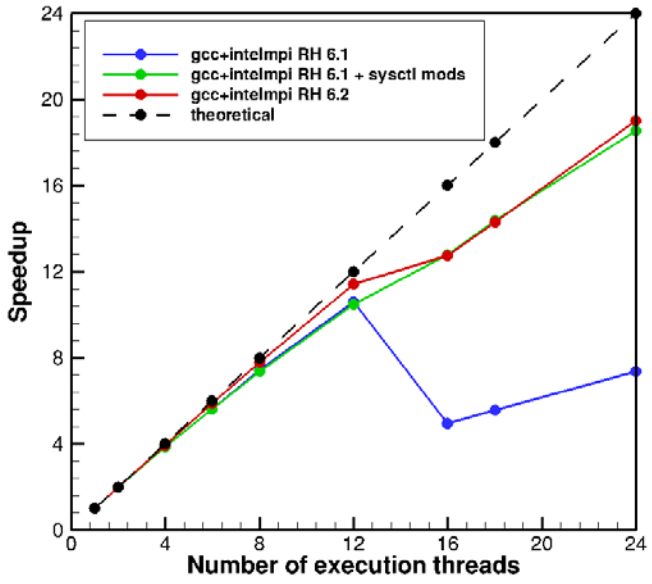

Figure 4. Comparison of application speedup using unpatched kernel with ASLR turned on (gcc+intelmpi RH 6.1) and off (gcc+intelmpi RH 6.1 + sysctl mods) and the patched kernel (gcc+intelmpi RH 6.2).

A small performance drop when using more than one core per compute unit can be observed in figure 4, but application performance is greatly improved over the unpatched kernel. This performance drop may be minimized in the future by further tuning for the Interlagos architecture with the fused multiply add instruction and optimizations for cache usage. An interesting discussion of the shared floating point unit can be found in the AMD Opteron 6200 Tuning guide (*7*).

> "The floating point unit is capable of producing four double-precision FLOPS per cycle per clock cycle simultaneously to each core in a pair for a total of eight per core pair per cycle. This is comparable to the floating point performance per core per cycle of an AMD Opteron™ 6100 CPU. But, unlike prior CPUs, when one core is issuing fewer floating point instructions, the other core in the pair can use its four FLOPS/cycle plus any unused by the other core to fully exploit the capacity of the Flex FP. For example, in the extreme case of one core executing no floating point instructions, the other core of the pair could achieve up to 8 double precision floating point operations per cycle."

The nonuniform memory access (NUMA) configuration of the Thufir compute nodes is given with the numactl command:

```
 [bjhenz@thufirc-0001 ~]$ numactl --hardware
available: 4 nodes (0-3)
node 0 cpus: 0 1 2 3 4 5
node 0 size: 16382 MB
node 0 free: 15584 MB
node 1 cpus: 6 7 8 9 10 11
node 1 size: 16384 MB
node 1 free: 15627 MB
node 2 cpus: 12 13 14 15 16 17
node 2 size: 16384 MB
node 2 free: 15622 MB
node 3 cpus: 18 19 20 21 22 23
node 3 size: 16384 MB
node 3 free: 15772 MB
node distances:
node   0   1   2   3
  0:  10  16  16  16
  1:  16  10  16  16
  2:  16  16  10  16
  3:  16  16  16  10
```

where a NUMA node is shown in figure 1a with eight cores as opposed to the six cores available in the Thufir system. The node distances reported for the adjacent nodes in the compute nodes without GPUs is 16, or, to put this another way, the cost to access data located on an adjacent node is about 60% greater than accessing local memory. As shown in the following example, this cost increases to about 100%, increasing from a distance of 16 to 20 for graphics nodes that contain GPUs.

```
[bjhenz@thufirg-0001 ~]$ numactl --hardware
…
node distances:
node   0   1   2   3
  0:  10  20  20  20
  1:  20  10  20  20
  2:  20  20  10  20
  3:  20  20  20  10
```

The network fabric of this system is based on 10 GigE Ethernet, as opposed to other high-performance network interconnects such as Infiniband. This choice of interconnect is based on the primary use of this system, namely emulation of mobile ad-hoc networks that require Ethernet for the transport of emulated OTA network traffic. Secondary applications used on the DHPI include GPU-based RF path loss calculations, network simulation, and other physics-based simulations that use MPI for message passing. MPI over Ethernet typically involves the

Transmission Control Protocol (TCP)/IP stack, increasing overhead, including CPU utilization and poorer performance in general. For this reason, hardware vendors have developed their own methods to reduce the operating system (OS) overhead.

A number of hardware vendors have developed remote direct memory access (RDMA) methods that bypass the OS stack and system drivers to improve latency and network throughput. Examples include iWarp from Intel, RoCE from Mellanox, and MX from Myricom. Thufir is configured with Mellanox ConnectX EN adapter boards for 10 GigE connectivity. RoCE is installed, but as of this writing, we have been unsuccessful in using RoCE over the Gnodal switches for simultaneous communications between more than two nodes. The use of MPI over TCP works well, but performance is expected to improve once the RDMA implementation is fully usable (*8*).

3.6 Final Acceptance and Entry Into Production

We discovered several important issues throughout final acceptance testing and system entry to production that may have been identified earlier through more thorough acceptance testing. Our focus throughout development of the acceptance test plan and acceptance testing has been on proper application execution and system interactions with a secondary focus on achieving maximum system performance. After acceptance, we identified some of the performance issues discussed previously that have prevented 100% system utilization, including RDMA over Ethernet, compilers, operating systems, and general system tweaking. We recommend the possible inclusion of some stressful performance metrics during future acceptance testing. However, the inclusion of such metrics will increase the cost and time of system acceptance. It could potentially limit the amount of hardware that could be purchased and ultimately affect system performance because additional funds must be allocated to testing. This process has been a delicate balancing act between the budget, application requirements, and the amount of system support required.

Large scale-MANET emulations have been performed on the system during testing and initial production, including an emulation with 5000 wireless nodes. Applications available for the GPUs, including RF propagation path loss using the ITM and ray tracing algorithms, have executed successfully with expected performance achieved. The GPU applications are primarily independent of the network interconnects and the MANET emulation uses TCP/IP for communications between nodes. Performance issues have been limited to secondary applications using the cluster that require high-performance MPI implementations. In addition to initially poor process thread placement on the CPU core, the RDMA over Ethernet implementation installed on the system has failed to perform across more than two nodes simultaneously. These issues have been solved through either OS and kernel updates or, as in the case of RDMA over Ethernet, are being resolved through the device vendor support.

4. Conclusions

This report details many of the issues encountered during installation and acceptance of the CPU/GPU hybrid cluster Thufir located at the U.S. Army Research Laboratory, APG, MD. The uniqueness of this system lies in the use of the low latency 10 GigE network fabric, many-core (12) AMD Opteron CPUs, Nvidia Fermi GPUs, and a mixture of CPU-only and CPU/GPU nodes.

5. References

1. Herbert, F. *Dune*. Chilton Books: Philadelphia, PA, 1965.
2. Rick, T.; Mathur, R. Fast Edge-Diffraction-Based Radio Wave Propagation Model for Graphics Hardware. *Proceedings of 2nd International ITG Conference on Antennas (INICA),* Munich, Germany, 28–30 March 2007.
3. Damosso, E. Ed., *COST Action 231: Digital Mobile Radio Towards Future Generation Systems, Final Report.* Office for Official Publications of the European Communities: Luxembourg, 1999.
4. Wang, F.; Sarabandi, K. A Physics-Based Statistical Model for Wave Propagation Through Foliage. *IEEE Transactions on Antennas and Propagation* **2007,** *55* (3).
5. Meyer, R. A.; Boyle, S.; Ollerton, B.; McKeon, D. Network Simulation of the Electronic Battlefield. *Proceedings of the Users Group Conference*, Williamsburg, VA, 7–11 June 2004.
6. Shared Level-1 Instruction-Cache Performance on AMD Family 15h CPUs. Advanced Micro Devices, Inc., White Paper, December 2011, http://www.naic.edu/~phil/software/amd/51803A_OpteronLinuxTuningGuide_SCREEN.pdf (accessed 26 August 2014).
7. *AMD Opteron 6200 Series Processors Linux Tuning Guide,* Advanced Micro Devices, Inc., April 2012, v1. http://amd-dev.wpengine.netdna-cdn.com/wordpress/media/2012/10/51803A_OpteronLinuxTuningGuide_SCREEN.pdf (accessed 26August 2014).
8. Gaiser, L.; Kraus, B.; Wernicke, J. Implementation and Comparison of RDMA Over Ethernet. LANL Institutes Information Science and Technology Institute, http:// institute .lanl.gov/isti/summer-school/cluster_network/projects-2010/Team_CYAN_Implementation_and_Comparison_of_RDMA_Over_Ethernet_Presentation.pdf, LA-UR 10-05188, (accessed August 26, 2014).

List of Symbols, Abbreviations, and Acronyms

ASLR	address-space layout randomization
APG	Aberdeen Proving Ground
ARL	U.S. Army Research Laboratory
CPU	central processing unit
CUDA	compute unified device architecture
DHPI	dedicated high performance computing project investment
DOD	Department of Defense
EMANE	extendable mobile ad-hoc network emulator
GigE	gigabit Ethernet
GPGPU	general purpose graphics processing unit
GPU	graphics processing unit
HPC	high performance computing
IP	Internet Protocol
ITM	Irregular Terrain Model
MANET	mobile ad-hoc network
MNMI	Mobile Network Modeling Institute
MPI	Message Passing Interface
NUMA	nonuniform memory access
OS	operation system
OTA	over the air
PCB	printed circuit board
RDMA	remote direct memory access
RF	radio frequency
SIMD	single instruction multiple data

SRW	Soldier Radio Waveform
TCP	Transmission Control Protocol
TI	technology insertion
VM	virtual machine
WNW	Wideband Networking Waveform

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

2 (PDF) DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL IMAL HRA MAIL & RECORDS MGMT

1 (PDF) GOVT PRINTG OFC A MALHOTRA

6 (PDF) DIR USARL RDRL CIH C B HENZ J KNAP RDRL CIH M N LAZORISAK J LIVINGSTON RDRL CIH S S PARK D SHIRES